# Week 18 : Distributions of Sample Statistics

| |
|---|
| pending tasks |
| type |

## Introduction - Inferential Statistics

- The focus of this module is **Inferential statistics.**
- Descriptive statistics is used to describe the attributes of the given data while inferential statistics is used to draw inferences based on the data, these cannot be directly interpreted from the data.
- The goal is to **infer something about the population parameters, given the visibility of sample measures.** It enables formulation of hypothesis backed by data.
- Distribution of sample statics involves studying the distribution of random variables (RVs). In inferential statistics, the distribution of RVs defined will be studied closely. (RVs are the sample statics).
- A sample is a subset of a population and a statistic is can be any of the measures of centrality/spread.


Fig.1 Foundational blocks of DS.

Foundational learning is quintessential for further progress in the course.

## Distribution of Sample Statistics

- The process of sampling should be seen as a random process and the sample statistics should be seen as random variables. In this sense statistics and probability are related in data science.
- The distribution of sample statistics should be related to the parameters of the population .i.e. drawing inferences about the population from the distributions of sample statistics.

**RECAP - Population**

- A population is a collection of a large number of **people, objects, or events** under study.
- e.g: all people in a country(people), a qualified subset of the population writing the 10th board exams at Bengaluru in 2021, all model S cars manufactured by Tesla(objects), detection of particles in the LHC(events).
- These can also be **hypothetical items .** e.g: all possible hands in a game of cards i.e 7 cards out of 52, all possible folds of a protein structure.

## Parameter

- A parameter is a numeric property of the entire population.
- E.g: avg. per capita income of all people in India (central tendency of mean), proportional outcome of an exam i.e proportion of students who passing the exam, the standard deviation of mileage across cars, average time interval (mean) since previous detection in LHC, proportion of hands with the king of diamonds, average effectiveness of a protein as a drug.
- It is not always possible (pragmatic) to calculate population parameters as the **population size** might be very large, the **data collection is expensive** or the **random process is unknown**. Thus, samples are created and studied.

## Sample

- A sample is a subset chosen from a population using a defined procedure. It is done only when the samples drawn are of interest. Multiple samples can be defined for a given population. The samples may intersect i.e. have common items from the population.

- E.g: All people in India in a given age-group, sampling students giving the board exam based on the schools, all cars manufactured in different factories or random groups of 1% of the cars manufactured (random sampling), all detections in LHC noted by a scientist, hands obtained on doing n shuffles.

**Statistic**

- A statistic is a numerical property of the entire sample.
- E.g: the mean income for each sample, proportion of students passing in the exam for a given school, the stddev of mileage across cars of the random samples, proportion of hands with a King of Diamonds in hands obtained from n random shuffles,

## Why do we Compute Statistics?

1. To describe the sample. It compactly represents the sample.

   E.g: the quick summary of school performance can be drawn from the summary statistics of pass percentage, the factory with least variation i.e. stddev in the car performance can be determined. Apart from single statistic the distributions can be visualised, compute statistics for sub-sets of the sample (i.e. consider a different attribute within the sample for analysis).

2. To estimate population parameter given the sample statics.
3. To test hypothesis about the population derived from the sample.

- Estimating population parameters and testing hypothesis are part of inferential statistics.
- The estimations methods used are mathematical relations that are  generally true and not bound to the dataset. The generality is based on two key assumptions.

## Estimate Population Parameters

*assumption 1 -  ( t = 1:15)*

- There are two key assumptions made, the first one relates only to the population and the second one to the processing of sampling.
- **Assumption 1 :** The values of interest of the elements in the population are independent random variables with a common distribution.
- The assumption is that such a curve exists and is common for all. Each element value in the sequence of values is independent and all follow the same distribution.

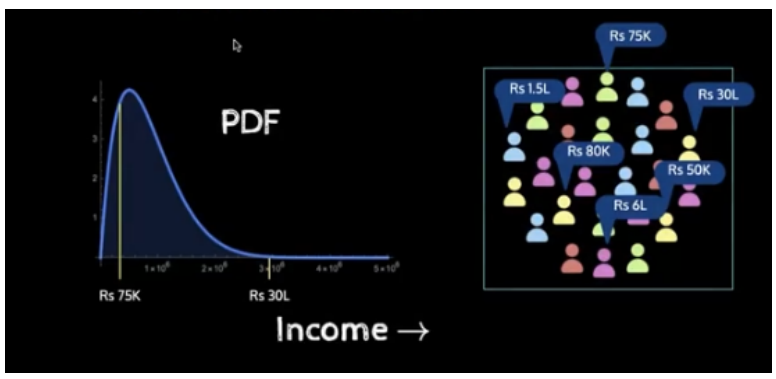



Fig.2a Example PDF for income in India .

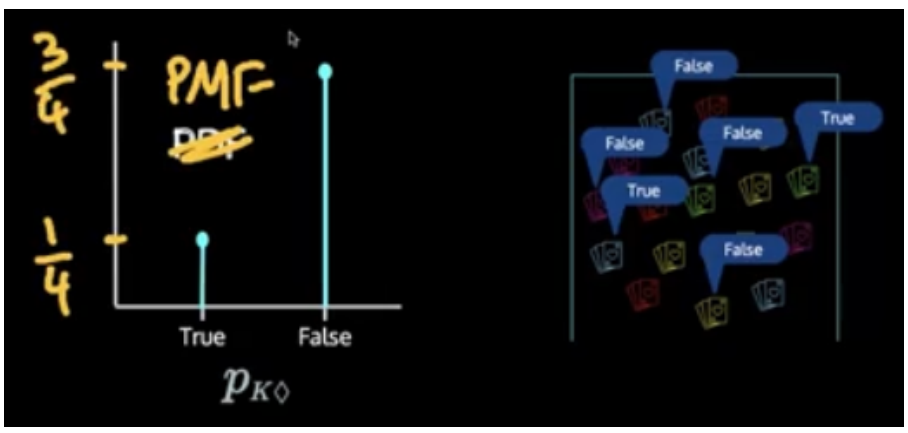



Fig.2b PMF of finding k-diamonds,

## Random Sample

*assumption 2 - ( t = 0:52 )*

- **Assumption 2: Each element** of the population has an equal chance of being selected in **any sample**.
- This is a constraint on the methods used to generate the random samples.
- E.g: The sampling method of using age group to split the population violates the assumption. This implies that the sampling method cannot be used for inferential statistics but can be used for descriptive statistics. Similarly, splitting the students based on schools violates the second assumption. Sampling cars by factory cannot be used but sampling random 1% of the cars satisfies the assumption. Here, note that the attributes of the cars are not used.

- This can be formalised by considering sampling as a random process where the outcomes are coming from a probability space.

## Recap : Probability

*def probability space - ( t = 0:58)*

*def random variable - ( t = 1:37)*

- A **probability space** is given by the triple $(\Omega, F, P)$ where, $\Omega$ is the set of possible outcomes, $F$ is the set of events, and $P$ maps events to probabilities.

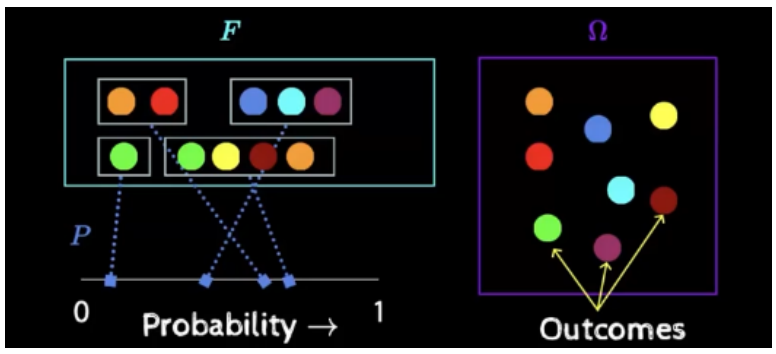



Fig. 3a Probability space triple, example representation.

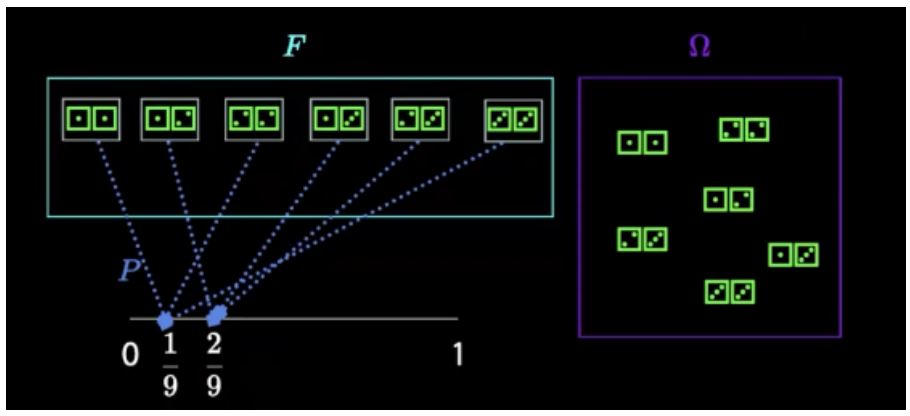



Fig.3b Example of three faced die tossed twice.

- A **random variable** is a **measurable function** that maps outcomes in a probability space to real numbers.

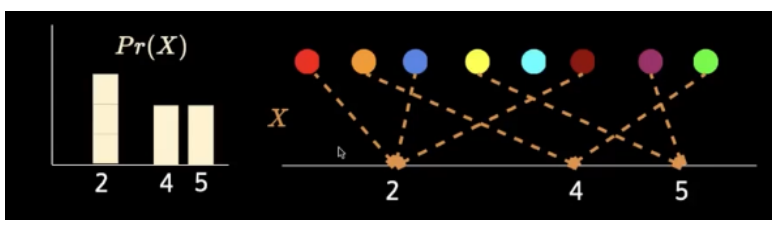



Fig.4a Example of RV X and it's distribution on left.

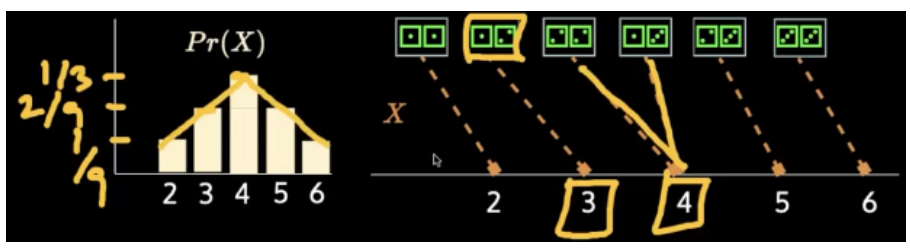



Fig.4c Example of X:sum of outcomes on two rolls of three faced die.

## Probability Space

- The concept of random sampling can be mapped to the probability space.
- $\Omega$ is the set of all possible samples that can be obtained through random sampling strategy. Since each element of the population should be equally likely to occur in each sample,  the following constraints must be satisfied:
  1. All  samples are of same size.
  2. All possible samples are outcomes.
  3. Probability of each sample is equal.

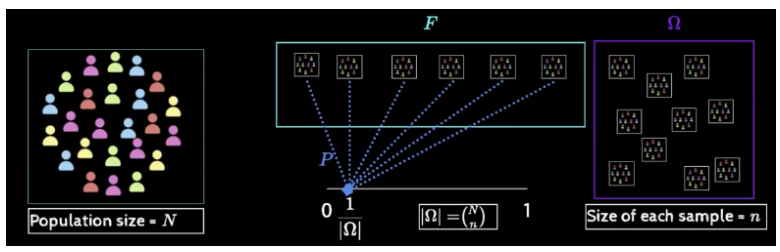



Fig.5a Example of mapping sampling to a probability space.

## What kind of Random Variables?

- Statistics can be used to map the samples to real numbers. Thus, statics can be used as RVs. If the income of population of N people is given, then the samples drawn from the population can be mapped to real numbers using the statistic mean. The probability distribution of the random variable can then be plotted.

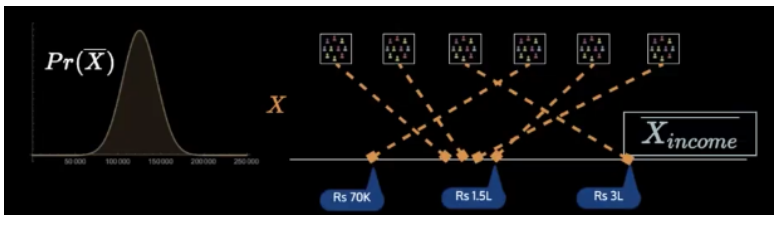



Fig.5b Mapping the samples to real values using statistic as RV and drawing PDF.

- Stddev and relative proportions can also be used as RVs.

## What is Inferential Statistics?

- Given the distribution of sample statistics, conclusions can be drawn on the population parameters using inferential statistics.
- To get an insight into the relationship between the parameters and distribution of sample statistics, we assume that population parameters are known and the distribution of RV(sample statistics) has to be drawn from this. The Central Limit Theorem is used for sample mean and variance and chi squared distribution is used for stddev and variance.

## Our Roadmap

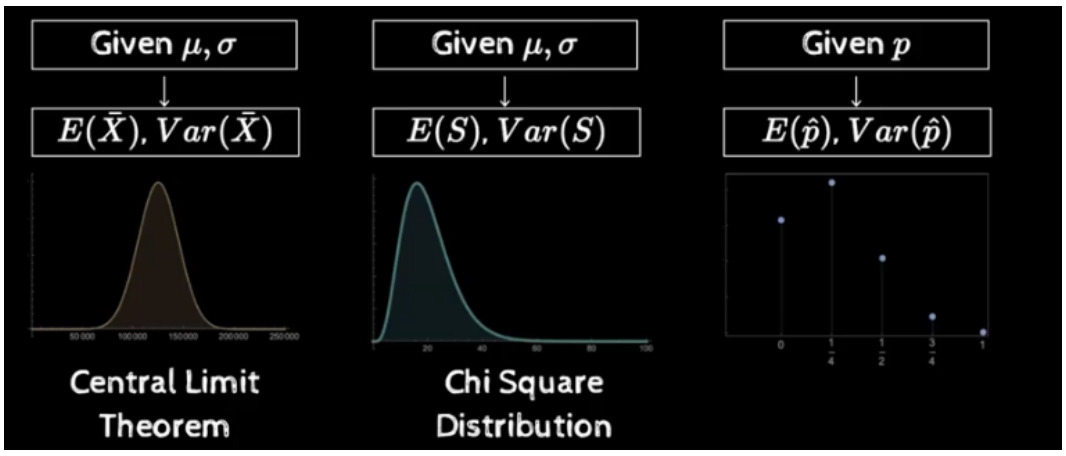



Fig.6 Roadmap for upcoming topics.

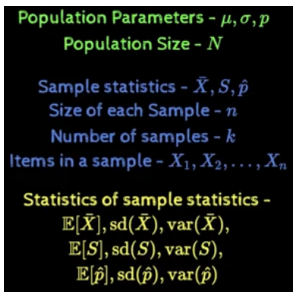



Fig.7 Notations

- The expected value of sample should be as close to the population parameter as possible. This implies, $\mathbb{E}[\bar{X}]$ should be close to $\mu$ .
- In the derivation, it can be seen that the expected value of each element of the sample is equal to the expectation of the population since, all events are assumed to be independent and equally likely.

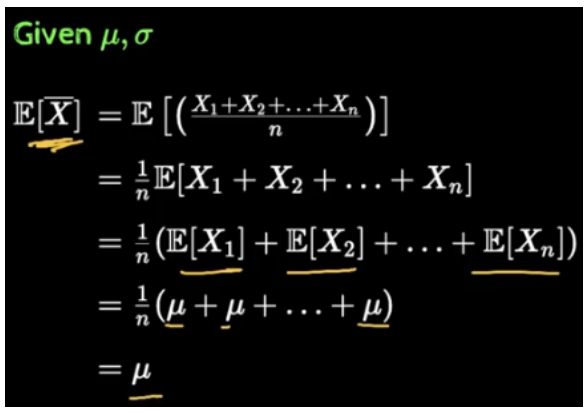



Fig. 8a Computing sample mean.

- The result is independent of the distribution $\sigma$ and the sample size n. $\bar{X}$ is an **unbiased estimate** of $\mu$ .i.e for any particular sample $\bar{X}$ can be less than or greater than $\mu$ but the average value is equal to $\mu$.
- Bias is a systematic error. which means the value is underestimated or overestimated.

## Demo 01

*def seed - (t = 7:50)*

- The mean for rolling three dies is discussed through a demo using Mathematica.
- It is observed the value approaches the actual mean with increase in number of samples(k) and converges faster(lower deviation, narrower distribution) when a larger sample size(n) is used.
- A seed is used to introduce some amount of determinism to the experiment.

## Demo 02

- The continuous distribution, the normal distribution was used to observe the previous outcomes of the premise.

## Demo Problems


Fig.9a,b Summary of previous demos.

- The previous two demos focused on providing an intuition for the fact that $\mathrm{E}[\bar{X}] = \mu$ .
- $Var(\bar{X})$ tells how much the individual $\bar{X}$ differ from the mean $\mu$. The ideal value is expected to be close to $0$. Also, it should not be greater than $\sigma^2$, which is the variance of the population.
- So far, random samples are taken from the population that satisfy the two assumptions. The sample statistic $\bar{X}$is calculated (instead of calculating the population parameter). The variance of the statistic $var(\bar{X})$ is calculated. variance thus calculated should not be more than the the population parameter $\sigma^2$.

## Exercise - Part1

**Exercise 1 :**For increasing values of sample size 'n', the variance of $\bar{X}$ over a 1000 samples was plotted.

- It was observed that the variance decreased with increasing n i.e are inversely proportional.

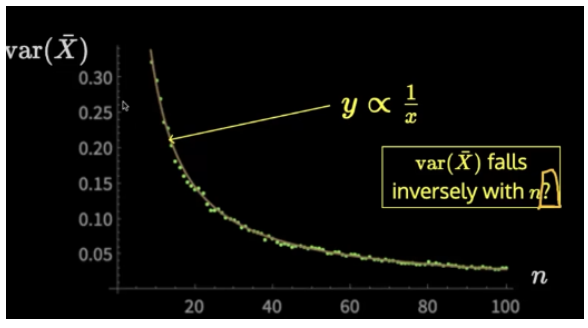
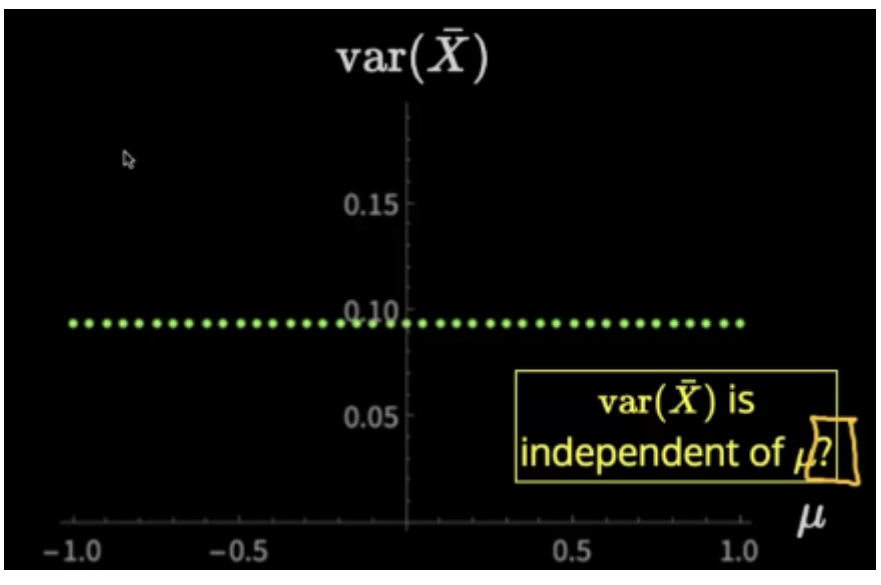



Fig.10a Variance is inversely proportional to n.

**Exercise 2:** Plot variance vs $\mu \; and \; \sigma$ for average of 10 numbers from a normal distribution over a 1000 samples.

- It is observed that $var(\bar{X})$ is independent of $\mu$ and it scales linearly with $\sigma^2$.



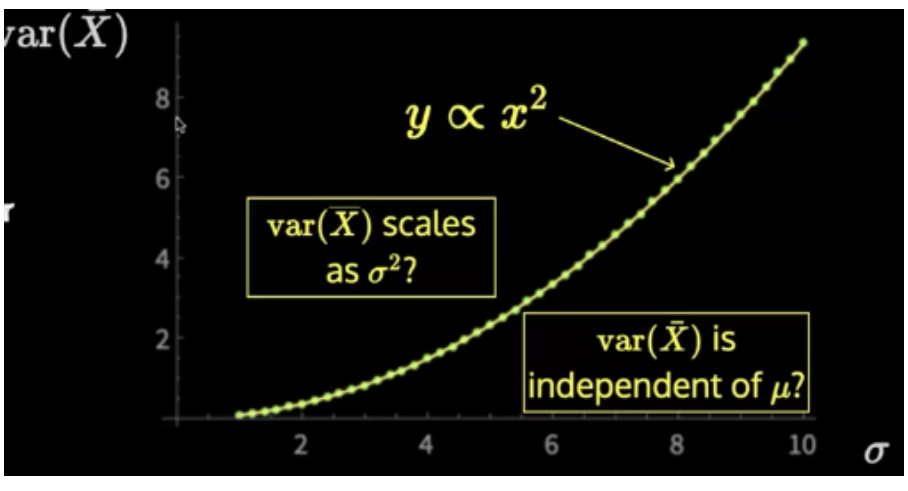




Fig.10 b,c Relationship of variance in samples and population parameters.

thus, $var(\bar{X}) = \frac{\sigma^2}{n}$.

- As already seen , $var(\bar{X}) = E[\bar{X}^2] - E[\bar{X}]^2$.
- $var(a\bar{X}) = a^2 var(\bar{X})$
- $var(X_1 + X_2) = var(X_1) + var(X_2)$

## Exercise - Part 2

- The derivation for the relation of variance to population parameters previously observed was derived.

$$
\begin{aligned}
\text{var}(\bar{X}) &= \mathbb{E}[\bar{X}^2] - \mathbb{E}[\bar{X}]^2 \\
\text{var}(a\bar{X}) &= a^2(\text{var}(\bar{X})) \\
\text{var}(X_1 + X_2) &= \text{var}(X_1) + \text{var}(X_2) \\
\text{var}(\bar{X}) &= \text{var}(\tfrac{X_1+X_2+\ldots+X_n}{n}) \\
&= \tfrac{1}{n^2}\text{var}(X_1 + X_2 + \ldots + X_n) \\
&= \tfrac{1}{n^2}(\text{var}(X_1) + \text{var}(X_2) + \ldots + \text{var}(X_n)) \\
&= \tfrac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) = \boxed{\tfrac{\sigma^2}{n}}
\end{aligned}
$$

- There is no dependence on the mean of the population.
- Taking a square root on both the sides gives the standard deviation:

$$sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- This implies, $\bar{X}$ is an unbiased estimate. But it varies from one sample to another with $sd = \frac{\sigma}{\sqrt{n}}$. For practical purposes it is better to think in terms of sd as it has the same unit as the mean.
- The desirable result is to have a low sd, this can only be achieved by increasing $n$ as $\sigma$ is a population parameter and it cannot be changed.
- E.g: to halve the sd, n has to be increased by 4.

## Summary

- Inferences about the population can be drawn given the sample statics.
- Sampling is a random process and the sample statistics are seen as RVs, thus, relating probability to statistics.
- Statistics are computed to compress the details in a sample. With regard to inferential statistics population parameters are estimated from given sample statistics, also, hypothesis about the population can be derived from the sample.
- There are two assumptions made that enable the estimation of population parameters from sample statistics:

  **Assumption 1 :** The values of interest of the elements in the population are independent random variables with a common distribution.

  **Assumption 2: Each element** of the population has an equal chance of being selected in **any sample**.
- For a population with mean $\mu$, it is seen that $\mathrm{E}[\bar{X}] = \mu$ , where $E[\bar{X}]$ is the expectation of the mean of samples.
- Also, $var(\bar{X}) = \frac{\sigma^2}{n}$ and $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.For practical purposes it is better to think in terms of sd as it has the same unit as the mean.